On a Meaningful Integration of Web Services in Data-Intensive Biomedical Environments: The DICODE Approach

Guillermo de la Calle[1], Miguel García-Remesal[1], Manolis Tzagarakis[2,3], Spyros Christodoulou[2,3], Georgia Tsiliki[4], Nikos Karacapilidis[2,3]

[1]*Biomedical Informatics Group, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain*

[2]*University of Patras, 26504 Rio Patras,Greece*

[3]*Computer Technology Institute and Press "Diophantus", N. Kazantzaki str., University of Patras Campus,26504 Rio Patras, Greece*

[4]*Bioinformatics and Medical Informatics Group, Biomedical Research Foundation, Academy of Athens, 115 27 Athens, Greece*

{gcalle, mgarcia}@infomed.dia.fi.upm.es, tzagara@upatras.gr, shristod@cti.gr, gtsiliki@bioacademy.gr, nikos@mech.upatras.gr

Abstract

This paper reports on an innovative approach that aims to reduce information management costs in data-intensive and cognitively-complex biomedical environments. Recognizing the importance of prominent high-performance computing paradigms and large data processing technologies as well as collaboration support systems to remedy data-intensive issues, it adopts a hybrid approach by building on the synergy of these technologies. The proposed approach provides innovative Web-based workbenches that integrate and orchestrate a set of interoperable services that reduce the data-intensiveness and complexity overload at critical decision points to a manageable level, thus permitting stakeholders to be more productive and concentrate on creative activities.

Keywords: Data-intensiveness, Machine intelligence, Human intelligence, Service integration, Information management.

1. Introduction

An enormous amount of content already exists today in the digital universe which is characterized by high rates of new information that is being distributed and demands attention. This enables professionals in various fields to have instant access to a wealth of information in order to successfully tackle their tasks at hand. Fields such as clinico-genomic research – investigating how genes are related to diseases– or marketing –aiming to forage the Web (blogs, forums, wikis, etc.) for high-level knowledge such as public opinions about its products and services– rely heavily on the wealth of available information. Yet, this wealth of information also poses immense problems as it can negatively affect the effectiveness of decision making in an organization [1] and create stress and cognitive overload to its stakeholders [2]. Professionals in such fields have to cope with data-intensiveness by using tools to appropriately assemble and analyze the enormous volumes of complex multi-faceted data from different sources. In many settings, they must also efficiently and effectively collaborate to make sense of these data and make the appropriate decisions.

As widely admitted, the pathologies of "big data" are primarily those of analysis [3]. Information management costs are important in fields of activities that rely heavily on support from machine intelligence (that include data mining and cloud computing infrastructures), as well as on support for augmenting collective intelligence (such as collaborative systems). Solutions to address information management tasks in such contexts have to face two major imperatives: (i) exploitation of the information growth by ensuring a flexible and scalable information infrastructure, and (ii) *e*xploitation of the competences of all stakeholders to meaningfully confront various information management issues. In other words, dealing with data-intensive and cognitively complex settings is not a technical problem alone.

This paper presents an innovative Web-based approach in the context of the Dicode project (http://dicode-project.eu/). The project adopts a hybrid approach building on the synergy between machine and human (collective) intelligence to facilitate sense and decision making in data-intensive and cognitively-complex environments. A common framework integrates a variety of services, tools and applications, allowing users to access a wide spectrum of resources through a single platform.

2. Background

Dicode project involves two main aspects: *data mining* and *collaboration and decision making support*. New challenges related to data storage, organization and searching have been established in modern Data Mining, where statistical problems exploded both in size and complexity [4]. Topics such as text mining [5] dealing with large collections of texts taken from big biological datasets or social media, or subgroup detection and data analysis processes for detection of local patterns in data are particularly relevant in the context of the Dicode project. This issue becomes interesting for the project since: (i) it is not possible to scale up algorithms by the straight-forward approach of sampling, i.e. always the complete data set has to be processed for pattern discovery, and (ii) supporting collaboration and decision making support by data mining is more adequately solved by tools that can present different facets of a problem, instead of a global rule that explains everything.

The success of these data mining solutions depends on external factors such as: (i) the ability to integrate the data mining solutions into the application environment, (ii) enabling the user and domain expert to guide and control the data mining process and include their domain knowledge, and (iii) the overall usability of the data mining system, in particular the ability to re-use existing solutions and built upon proven solutions.

On the other hand, the term "Collaboration Support Systems" describes those applications designed to assist a group of people working in a common task to achieve their objectives. The emergence of the Web 2.0 era introduced a great range of collaboration tools which provide engagement at a massive scale. The Dicode project presents a novel approach that allows the integration of different collaboration and decision making support tools regarding argumentative collaboration, sharing, note taking and mind mapping under a common framework.

3. The DICODE approach

The Dicode approach aims at ensuring a seamless integration and interoperability of services, not only from a technical but also from a conceptual point of view. In this regard, semantics techniques have been thoroughly used to define an ontological framework for capturing and representing stakeholder perspectives [7] towards augmenting collaboration and decision making.

We have designed and implemented the Dicode workbench, a web application which provides end-users with an integrated environment of services. A web-based solution was adopted considering the benefits provided by this approach such as the platform and operating system independency, accessibility, availability, security issues and the simplicity and facility of use for the end-users. The Dicode workbench allows users to create different workspaces. Each workspace is customizable by end-users according to their own needs.

Accessibility to workspaces can be defined as public or private. Public means that all users in the system can access the workspace, while private workspaces can be accessed by a list of authorized users. A first functional prototype of the workbench is publicly available at http://hodgkin.dia.fi.upm.es:8080/dicode.

In Dicode, integration has been done at two levels: (i) at user interface (UI) level, and (ii) at operational level. Integration at UI level deals with the visualization of services to end-users. On the other hand, integration at operational level aims to ensure the communication and exchange of data between different services. Regarding to UI integration, a widget-based approach [8] has been adopted to deliver the Dicode services to end-users. Users can interact with the widget content but such content is not usually controlled by the widget host. Widget behaviour is directly managed by the widget itself. The widget host manages only visualization aspects, such the location or size of widgets within the host page.

Services already developed include scalable data mining, collaboration support and decision making support services. Figure 1 shows an example of a workspace within the Dicode workbench. Widgets are used to give end-users access to different types of services such as data acquisition and pre-processing services, data mining services, and collaboration and decision making support services. Separate widgets can interact and exchange data through a "drag & drop" functionality.

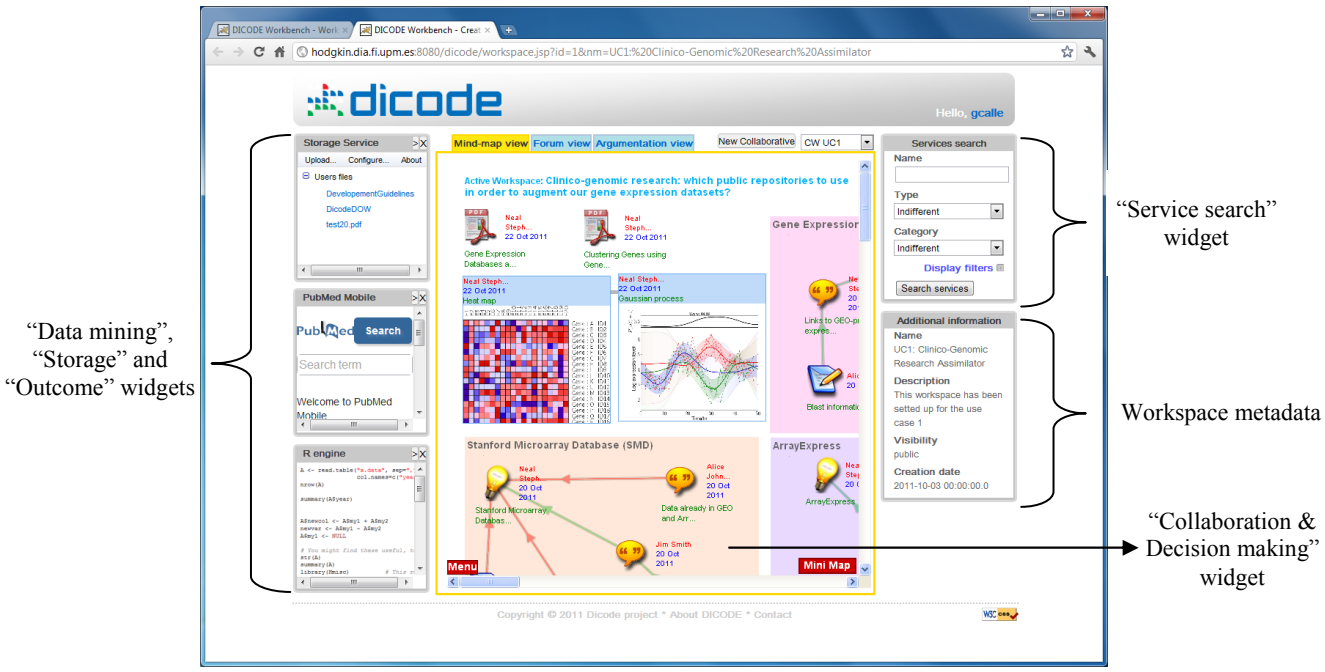

Figure 1. Screenshot of the Dicode workbench. The widget at the center provides access to collaboration support services. On the right, the search service widget allows users to locate new services. Widgets on the left are customizable by the users.

Regarding to operational integration level, we have developed the Dicode Registry of Services (DRS). DRS is a service accessible via REST services that keeps meta-data information related to all available services that include name of the service, service provider, its location and functionality. DRS is aimed to provide support to services for data mining, collaboration and decision making support, recommendations, and searches. Furthermore, DRS also contains service annotations using concepts of the Dicode ONtology (DON) [7]. These annotations are established by the service provider during the registration process of the service in the Dicode workbench. Such annotations are used, for instance, to facilitate users to locate services within the platform.

4. An example of use

In this section, we present an example scenario from the clinico-genomic research domain and discuss how the integrated workbench can address the types of problems that emerge in such environments.

Consider two researchers, Jim and Alice, aiming to investigate which genes or groups of genes are associated with breast cancer disease. Initially, they create a new collaboration workspace where they exchange ideas related to which data sources to use, based on their own data analysis experience and literature knowledge. They search relevant literature via PubMed using the appropriate search services.

Jim conducts a preliminary analysis with some in-house gene-expression datasets; however, his findings were not very encouraging, which was attributed to the small sample size available. Jim and Alice discuss about how to overcome that problem, and they decide to augment their samples with publicly available gene-expression data sources. After deciding what data to use, they keep collaborating in order to discuss how the data will be processed. They decide to first analyze the datasets using the well-known Significant Analysis of Microarrays (SAM) methodology [9]. Jim is also offering to provide all the necessary R scripts (http://www.r-project.org/) for this initial statistical analysis. In addition, they decide to employ model-based data integration methodologies [10-12] that have been recently published and claim to perform better than simple data integration techniques [13]. Both researchers can execute the available services, retrieve the results, and interpret them in terms of the initial research question.

The above scenario gives an indication at the range of tools that clinico-genomic researchers nowadays use in their daily work practise. Jim and Alice require tools to search online data repositories and store useful information, tools to facilitate their collaboration permitting sharing and sense-making of available

resources, as well as tools to meaningful analyse these resources via elaborated algorithms and discuss the outcomes. Although a number of tools are available to support each need, in the context of Jim's and Alice's work these tools can only be used separately and as standalone applications as they provide limited or no support for information exchange. In the context of today's available tools, information management related tasks —as outlined in the above scenario— cannot be streamlined and automated, which leads to increased information management costs that obstruct the actual work. In general, today's tools do not integrate well into work practices that rely equally on support for machine and human reasoning.

The Dicode workbench has been conceived to alleviate such problems by enabling the smooth integration of all required tools into Jim's and Alice's daily work practice. In particular, using the Dicode approach, Jim and Alice can create a new workspace and customize it by selecting the necessary services to support their research goal. Using the Collaboration widget, they can share and discuss the resources in order to assess their usefulness. Data-mining widgets allow Jim and Alice uploading scripts they program and execute them to process resources. Furthermore, they can also consult public online database such as PubMed from the same workspace.

The Dicode workbench not only makes a range of services readily available to researchers; since it keeps together datasets, algorithms and their outcomes along with all collaboration sessions, it functions as an archiving platform where the provenance of results can be maintained and contextualized, thus further facilitating sense-making.

5. Conclusions

Building on current advancements, the approach presented in this paper exploits and builds on the synergy between the reasoning capabilities of the machine and humans and provides an integrated environment for knowledge workers to cope with data-intensive situations. The innovative workbenches available in Dicode incorporate and orchestrate a set of interoperable services that reduce the data-intensiveness and complexity overload at critical decision points to a manageable level, thus permitting stakeholders to be more productive and concentrate on creative and innovative activities. The Dicode workbench has already integrated several services such as the collaborative service, the storage service and the PubMed searcher.

6. Acknowledgements

This publication has been produced in the context of the EU Collaborative Project "DICODE - Mastering Data-Intensive Collaboration and Decision Making", which is co-funded by the European Commission under the contract FP7-ICT-257184. This publication reflects only the authors' views and the Community is not liable for any use that may be made of the information contained therein.

7. References

[1] A. Schick, "Information overload: a temporal approach", *Accounting, Organizations and Society*, 1990, 15(3), pp. 199-220.

[2] D. Kirsh, "A Few Thoughts on Cognitive Overload", *Intellectica*, 2000, vol. 1, nº 30, pp. 19-51.

[3] Economist, *A special report on managing information: Data, data everywhere*, Economist, 2010.

[4] T. Hastie et al., *The Elements of Statistical Learning*. NY, Springer Verlag, 2002.

[5] R. Feldman and J. Sanger, *The text mining handbook*. Cambridge University Press, 2007.

[6] N. Karacapilidis and D. Papadias, "Computer Supported Argumentation and Collaborative Decision Making: The HERMES system", *Information Systems*, 2001, 26(4), pp. 259-277.

[7] D. Thakker et al., "Socio-technical ontology development for modelling sensemaking in heterogeneous domains", in *Workshop on Ontologies come of age in the Semantic Web (OCAS2011)*, 2011, vol. 809, pp. 60-71. Available: http://ceur-ws.org/Vol-809/paper-08.pdf.

[8] R.R. Swick and M.S. Ackerman, "The X toolkit: more bricks for building user interfaces, or widgets for hire", in *Usenix Winter 1988 conf.*, pp. 221-228. Available: http://s.niallkennedy.com/papers/xtk.pdf.

[9] V.G. Tusher et al., "Significance analysis of microarrays applied to the ionizing radiation response", in *Proc. Natl. Acad. Sci. USA*, 2001, 98(9), pp. 5116-5121.

[10] C. Huttenhower et al., "A scalable method for integration and functional analysis of multiple microarray", *Bioinformatics*, 2006, 22(23), pp. 2890-2897.

[11] E. Garret-Mayer et al., "Cross study validation and combined analysis of gene expression microarray data", *Biostatistics*, 2007, 9, pp. 333-354.

[12] A.A. Shabalin et al., "Merging two gene-expression studies via cross-platform normalization", *Bioinformatics*, 2008, 24(9), pp. 1154-1160.

[13] R. Mrowka et al., "Does mapping reveal correlation between gene expression and protein-protein interactions?", *Nat Genetics*, 2003, 33, pp. 15-16.